Oberauer, K., Lewandowsky, S., Awh, E., Brown, G. D. A., Conway, A., Cowan, N., Donkin, C., Farrell, S., Hitch, G. J., Hurlstone, M. J., Ma, W. J., Morey, C. C., Nee, D. E., Schweppe, J., Vergauwe, E., & Ward, G. (2018). Benchmarks provide common ground for model development: Reply to Logie (2018) and Vandierendonck (2018). *Psychological Bulletin*, *144*(9), 972-977. https://doi.org/10.1037/bul0000165

Peer reviewed version

Link to published version (if available):
10.1037/bul0000165

Title page with All Author Information

Benchmarks for Models of Short-Term and Working Memory –

Response to the Commentaries of Logie and Vandierendonck

Klaus Oberauer, University of Zurich

Stephan Lewandowsky, University of Bristol and University of Western Australia

Edward Awh, University of Chicago

Gordon D. A. Brown, University of Warwick

Andrew Conway, Claremont Graduate University

Nelson Cowan, University of Missouri

Christopher Donkin, University of New South Wales

Simon Farrell, University of Western Australia

Graham J. Hitch, University of York

Mark J. Hurlstone, University of Western Australia

Wei Ji Ma, New York University

Candice C. Morey, Cardiff University

Derek Evan Nee, Florida State University

Judith Schweppe, University of Erfurt

Evie Vergauwe, University of Geneva

Geoff Ward, University of Essex

The preparation of this article was facilitated by two grants from the Swiss National Science Foundation that funded two workshops (grant numbers IZ 32Z0 150896 and IZ 32Z0 160296).

Correspondence concerning this article should be addressed to Klaus Oberauer or Stephan Lewandowsky., E-mail: k.oberauer@psychologie.uzh.ch or stephan.lewandowsky@bristol.ac.uk

Benchmarks for Models of Short-Term and Working Memory –

Response to the Commentaries of Logie and Vandierendonck

Abstract

We respond to the comments of Logie and Vandierendonck to our article proposing benchmark findings for evaluating theories and models of short-term and working memory. The response focuses on the two main points of criticism: (1) Logie and Vandierendonck argue that the scope of the set of benchmarks is too narrow. We explain why findings on how working memory are used in complex cognition, findings on executive functions, and findings from neuropsychological case studies are currently not included in the benchmarks, and why findings with visual and spatial materials are less prevalent among them. (2) The critics question the usefulness of the benchmarks and their ratings for advancing theory development. We explain why selecting and rating benchmarks is important and justifiable, and acknowledge that the present selection and rating decisions are in need of continuous updating. The usefulness of the benchmarks of all ratings is also enhanced by our concomitant on-line posting of data for many of these benchmarks.

Key words: Working memory, short-term memory, executive functions, benchmarks, commentary

We appreciate our colleagues' thoughtful commentaries on our article (Oberauer et al., in press). They raised several important issues and pointed out limitations of our efforts to compile a set of benchmarks for theories and models of short-term and working memory. We take their arguments as constructive critiques that give us an opportunity to clarify what we intended to achieve with our proposal, and assist the further development of the present benchmarks towards a broader consensus among researchers about the empirical basis of our theoretical efforts. Here we respond to their arguments.

Is the Scope too Narrow?

Both commentators express concerns about our decision to exclude findings on the role of working memory (WM) in specific cognitive activities, such as arithmetic, reasoning, and language processing, as well as the relation of working memory to executive functions (Logie, in press; Vandierendonck, in press). Vandierendonck in particular would have preferred a set of benchmarks for unified theories of cognition. Building a unified theory of cognition is a different, much more ambitious endeavor than building a comprehensive theory of WM. We believe that there is value in working towards a better theory of WM in its own right. Such a theory will form a solid foundation for understanding how WM contributes to other aspects of our mental life. Eventually, the success of a theory of WM will depend on how useful it is for that purpose. However, aiming to explain with high priority the contributions of WM to the broad range of complex human cognitive activities would set the bar impossibly high for a theory of how WM works, given our limited state of knowledge.

The Use of Working Memory in Complex Cognition

Vandierendonck cites several lines of research on the role of WM in arithmetic, reasoning, and language comprehension. The predominant results from this research address which processes in arithmetic, reasoning, or language processing depend on WM. As such, they inform theories of

these cognitive activities, but they do not tell us much about WM itself, other than underscoring its importance within the cognitive system.

We agree that it is crucial for an adequate conceptualization of WM to consider its role in cognition more generally: What is it there for, and how does it accomplish its function? At the same time, we disagree with Logie's claim that asking how WM theories explain the use of WM in everyday life is a "stronger test" than focusing on the "artificial tasks" from which most of the benchmark findings arise (Logie, in press; p. 10). We are not aware of any instance in which research on the use of WM for a particular cognitive task or activity has been leveraged to adjudicate between competing theories of WM. This paucity of relevant instances is unsurprising: There are principled reasons for why making such an argument is difficult. Take, for instance, research on how the syntactic complexity of sentences affects the difficulty of sentence comprehension, and of concurrent maintenance of a WM load (e.g., Loncke, Desmet, Vandierendonck, & Hartsuiker, 2011). Any explanation of findings from this kind of study will have to make assumptions not only about WM but also about sentence comprehension, and in particular syntactic parsing. Therefore, any inference in favor or against an assumption about WM is likely to stand and fall with the assumptions about language processing that it is combined with to explain the data. There is thus a good reason why WM researchers often prefer simple, artificial tasks: Explaining findings from them does not require as many risky assumptions about how people process the task in addition to assumptions about WM.

Executive Functions

The relation of WM to executive functions (EF) deserves special consideration. Many theorists assume that the constructs WM and EF are closely related (e.g., Baddeley, 1986; Kane & Engle, 2002); others don't share that view (e.g., Just & Carpenter, 1992; Ma, Husain, & Bays, 2014). The matter is complicated by the fact that the scope of the term EF is not well defined itself. We were aware that any decision on how to draw the line between findings to consider and findings not to consider as benchmarks in this thematic area could be perceived as biased in one or the other direction. We decided to include only those findings that speak to the question how, and how

strongly, WM and EF are related, and to exclude those that speak to questions about how specific forms of EF operate in specific paradigms or phenomena (e.g., how conflict is resolved in the flanker task, or what role inhibition plays in retrieval-induced forgetting). Because the concept of EF is ill defined and strongly theory-dependent, we included findings on the relation between WM and EF without using the term EF, using instead more specific and arguably more precise terms. We identified the following lines of evidence on the relation between WM and EF:

(a) Some authors consider all forms of processing that involve central attention (e.g., response selection, retrieval from LTM) as instances of EF (e.g., Szmalec, Vandierendonck, & Kemps, 2005). On this definition, the effects of concurrent processing on maintenance (BM 5.1) are an instance of the interplay of WM and EF.

(b) Updating of WM has been identified as one form of EF (Miyake et al., 2000). Where available, we included evidence from experimental paradigms involving updating across all benchmarks (see task codes NB and MU in the reference tables of the target article). However, few findings that met the criteria for benchmarks were specific to the updating of WM. This could reflect the fact that, despite its prominent place in the conceptual network of EF, the process of updating WM has not yet attracted enough systematic efforts to establish robust, general, and theoretically informative findings (for some recent efforts in that direction see Ecker, Lewandowsky, Oberauer, & Chee, 2010; Kessler & Oberauer, 2014; Rac-Lubashevsky & Kessler, 2016). One could argue – as Vandierendonck does – that switching between task sets is an instance of updating (procedural) WM, and task-set switching research has amassed a wealth of well-established findings (reviewed by Vandierendonck et al., 2010). We decided against including this body of work because the interpretation of task-set switching as an instance of WM updating is, as far as we see, far from universally accepted. Unsurprisingly, therefore, theories and computational models of task-set switching are to the most part specific to task-set switching and do not speak to other aspects of WM.

(c) Cognitive control (i.e., avoiding distraction, avoiding strong but wrong action tendencies) is universally acknowledged as a prototypical EF. Cognitive control has been related to WM primarily through correlational findings, which figure as BM 12.6. Other research has investigated how cognitive-control demands affect concurrent maintenance of unrelated material in WM (Barrouillet, Portrat, & Camos, 2011; Liefooghe, Barrouillet, Vandierendonck, & Camos, 2008), and found that there is nothing special about control demands – they have the same effect on maintenance as other central processing demands. We therefore included these findings in BM 5.1. Hence, we have included cognitive control phenomena in the benchmarks, albeit not under a dedicated benchmark labeled "cognitive control".

Neuropsychological Case Studies

Logie criticizes the fact that we largely excluded neuropsychological case studies that demonstrate dissociations between functions of working memory. The problem with single-case studies is that it is difficult to establish their replicability, and even harder to establish their generality, because case studies are, by definition, not a representative sample from any population (for related arguments see Miyake, Carpenter, & Just, 1994; Miyake, Carpenter, & Just, 1995). At the same time, we acknowledge that neuropsychological case studies have been very important in informing theories of working memory (Baddeley, Gathercole, & Papagno, 1998; Della Sala, Logie, Trivelli, Cubelli, & Marchetti, 1998). We hope that, for a future update of the benchmarks, we will find a way to systematically establish which findings from neuropsychological case studies are sufficiently robust and general to warrant crediting them with benchmark status.

Findings with Visual and Spatial Materials

Vandierendonck further criticizes what he considers to be our selective omission of many findings with visual and spatial materials. Specifically, he proposes that the syllable-based word-length effect (BM 7) is just one example of a broader set of complexity effects that includes phenomena in the spatial domain such as effects of path length, path crossings, and symmetry. In

our discussions we considered this expansion of BM 7, but eventually decided against it because we thought that it would be an overly strong theoretical commitment to subsume this diverse set of effects under a common category of complexity effects. Doing so would have expressed a certain degree of conviction that these effects are likely to have a common explanation, and we were not sufficiently confident in that assumption. The commonality of these phenomena is less clear than in other cases where we did subsume similar phenomena under a general description, such as the effects of presentation time for visual arrays and those of presentation rate for verbal lists (BM 2.4). We are confident that future empirical work on the effects of the characteristics of spatial paths, spatial configurations, and also non-spatial visual configurations (Brady & Alvarez, 2011; Brady & Tenenbaum, 2013) will establish many of them well enough to include them in the next update of the benchmarks.

How to use Benchmarks to Evaluate Theories?

A second point of skepticism raised by both commentators pertains to how the benchmarks should be used to evaluate theories. Should a theory that explains N benchmark findings be preferred over one that explains only N-K benchmark findings but additionally explains K other findings that are not benchmarks? Should a theory that explains N benchmarks with A ratings be preferred over one that accounts for equally many benchmarks with B or C ratings?

If the theory is intended as a theory of WM, then our answer is Yes. The premise of our endeavor is that not all empirical findings are equally important for evaluating theories. A theory's ability to explain findings that are well-replicated, that generalize over a broad range of materials, experimental paradigms, and populations, and that have informed theoretical decisions and debates, should count for more than its ability to explain findings lacking one or more of these features. Our critics have not presented an argument against this position. There might be exceptions – it is conceivable that a finding, although being highly specific to one paradigm and one kind of material, has profound implications for theories, and as such should be regarded a benchmark. We are open to such an argument and would consider including such a finding in the next version of the benchmarks.

We hope that our article lays the ground for a conversation about the importance of empirical phenomena for theory development and theory testing, and that through such a conversation the set of benchmarks will continue to evolve.

Vandierendonck, if we understand him correctly, accepts our general premise but believes that our selection and rating of benchmarks is biased. At the same time, Logie argues that a democratic approach – a popularity contest among empirical findings – is not a suitable approach for prioritizing findings, and we agree. These two comments highlight the inevitable tension between judgments made by a small group of experts – with a high risk of individual biases – and judgments based on a much broader group of stakeholders – risking the irrationalities of a popularity contest. We tried to strike a balance between these two risks by making selection and rating decisions through a consensus among researchers with diverse views[1], based on explicit criteria, and informed – but not determined – by a broader expert survey. We maintain that the result is much less biased than the selection of findings undertaken by any individual theorist or team. We propose the benchmarks in the target article as a first step in the direction of a systematic, rational, and unbiased ranking of findings by their importance for theory building and evaluation – not as its end product.

Logie takes issue with the ranking of benchmarks as A, B, or C, arguing that the A-ranked findings are not the ones that are most important for theory evaluation, but those that are most popular among researchers because they are the easiest to research, or the easiest to explain, or the ones that have been around long enough to be replicated and extended many times. As the present collection of benchmarks is a snapshot at one historical time point, it inevitably reflects historical trends in our discipline: There are more findings with verbal than visual and spatial materials in part because, for a long time, it was easier to present verbal stimuli to participants. Serial-position curves (an A benchmark) are more robustly and more generally established than the neurally silent short-

[1] Some of us propose unitary theories of working memories whereas others advocate multi-component theories; some of us believe that decay plays an important role in explaining short-term forgetting, others question such a role, and others still assume that working memory is limited by a discrete capacity, or a continuously varying resource. Some of us see a clear division between short-term/working memory on the one hand and long-term memory on the other, whereas others prefer a more unitary view of memory.

term maintenance of information (a C benchmark) because the former were first described more than a century ago (Nipher, 1878), whereas the latter had to await methods that were developed only very recently. There is no way to compensate for these imbalances except for researchers to fill in the missing evidence in the coming decades: Some of the novel findings will turn out to be replicable and general; others will not – until we know which of them do, it would be premature to assign them high evidential value for theory decisions.

If we grant WM researchers some degree of rationality in choosing their research questions, the historical trends are not entirely arbitrary: Findings that speak to important theoretical questions tend to attract follow-up research, and findings that are replicable and general tend to be built on by further empirical work that establishes them even more firmly. We assigned benchmarks rank A not merely on the grounds that they have been demonstrated very often, but on the grounds that they have attracted much research for good reasons.

We do not think that there is any risk that phenomena that missed out on the highest priority ranking for modelling will now be ignored. To the contrary, where we rated benchmarks as B or C because these phenomena required additional replication or generalization, that may even signal to researchers that a particular phenomenon is widely viewed as a fruitful target for new research efforts.

In other instances, findings were rated as B or C because they reflect more nuanced aspects of the more overarching category A benchmarks (for example, in serial recall fill-in and infill errors are a more nuanced feature of the locality constraint on transpositions; see BM 4.1.1 and 4.1.2). It is therefore reasonable to start with a theory that can explain the A benchmarks before accounting for the B and C benchmarks. We agree with Logie that the latter benchmarks may ultimately have more power to discriminate between theories, but focusing first on explaining the B and C benchmarks engenders the risk that theorists may invoke specialized mechanisms and assumptions that undermine a theory's ability to capture the A benchmark that describes the more general phenomenon. A model of the solar system has to explain seasons, the lengths of days and the

retrograde motion of Mars first, before we would worry about whether it can also predict a lunar eclipse. A model that predicts the latter but not the former is of little use, and no astronomer would seriously entertain it.

One final comment on the ratings: They are intended as guidelines for theorists aiming to explain how short-term or working memory works. Other theoretical aims warrant other selections and rank-orderings of phenomena to explain. In particular, a researcher who aims to build a theory with a narrower scope (e.g., focusing on the role of verbal serial recall in word learning) would be well advised to focus on those benchmarks that are most relevant to that endeavor, and give higher priority to B and C benchmarks within that set than to A-rated benchmarks outside it.

Are the Benchmarks Constraining Research?

Logie worries that the publication of a set of benchmarks discourages the search for novel findings. The past decades have witnessed a high rate of novel empirical discoveries, accompanied by a much slower rate of theory development. The strong incentives for generating novel empirical results – after all, they are the most common entrance ticket for peer-reviewed publications – will not go away soon. We are not worried that the empirical innovation in our field might slow down; rather, we are worried that it might continue through idle cycles of phenomena being discovered, firmly established, and analyzed in much detail by a flurry of studies until the empirical landscape becomes so complex that all hope for a complete explanation dissipates, upon which the area is abandoned. We believe that, as scientists, we have a responsibility to not only generate new findings but also to strive for better, more comprehensive explanations of the existing ones. Doing so will put more emphasis on developing strong theories which, in turn, would guide future empirical research towards questions of theoretical relevance, and away from phenomena-driven research.

Logie further argues that focusing our theoretical efforts on well-established findings is post-hoc and even circular, and that theory tests should focus on new predictions instead. This argument would have force if we lived in a world in which most competing theories can explain most of the

existing, well-established findings, so that we need to look for new empirical tests to adjudicate between them. As we see it, the reality of WM research is far from such a state. No existing theory or model of WM currently provides a joint explanation of even the A-rated benchmarks. As we elaborate in the Discussion of the target article, a satisfactory explanation would require a theory or model to not only be compatible with a finding but to imply it in such a way that the absence of that finding would be incompatible with the theory (Roberts & Pashler, 2000). Moreover, a joint explanation requires the theory or model to use the same assumptions, and ideally the same (or similar) parameter values for its explanation of each benchmark. This is a high standard for theorists, and a theory or model that reaches it would have considerably more explanatory power than any existing one. Therefore, the benchmarks are highly informative for adjudicating between theories.

Logie proposes an alternative approach to using empirical findings for evaluating theories. His proposal is to put together a set of well-established findings without differentiating among them by priority. Researchers should compare pairs of competing theories on the subset of findings that are associated to one or both of the competing theories. We fully agree with the rationale of this approach. The proposal is to select findings – among those for which replicability and some degree of generality has been established – by whether they are relevant for the theories in question. We understand our effort as an attempt to generalize the approach that Logie proposed, considering not just two competing theories but all currently competing theories of short-term and working memory jointly.[2] To that end we defined *theoretical leverage* – the ability of a finding to inform, and adjudicate between, existing theories of short-term and working memory – as a criterion for selecting and rating benchmarks.

[2] We also considered past theories that were eliminated from the competition by certain findings. For example, the mixed-list similarity effect, BM 8.1.2, was important to rule out chaining models of serial recall. These findings are included in the benchmarks because every new theory must explain them to not immediately meet the fate of those older theories.

Concluding Remarks

The process of putting together a set of benchmarks is arguably more important than the current outcome. The set we proposed in the target article certainly has its limitations, and will evolve over the coming decades. Our main purpose was to highlight the need for a shared set of targets for theories and models to explain, and to propose a process by which researchers in a field can work towards a consensus on these targets. We hope that others working on short-term and working memory will join us to carry this effort into the future, and that colleagues working in other fields of research will be inspired to embark in similar endeavors. To get involved, interested researchers could comment on the benchmarks on our web page,[3] or get in touch with one of the first two authors for proposing a new benchmark or a revision of the existing ones.[4] Researchers can also offer data sets for any existing or newly-accepted benchmarks to be included in our unified, on-line posting.[5]

[3] URL: https://wmbenchmarks.wordpress.com/
[4] k.oberauer@psychologie.uzh.ch; stephan.lewandowsky@bristol.ac.uk
[5] URL: https://github.com/oberauer/BenchmarksWM.git, and https://osf.io/g49c6/

References

Baddeley, A. D. (1986). *Working memory*. Oxford: Clarendon Press.

Baddeley, A. D., Gathercole, S. E., & Papagno, C. (1998). The phonological loop as a language learning device. *Psychological Review, 105*, 158-173.

Barrouillet, P., Portrat, S., & Camos, V. (2011). On the law relating processing to storage in working memory. *Psychological Review, 118*, 175-192.

Brady, T. F., & Alvarez, G. A. (2011). Hierarchical encoding in visual working memory: Ensemble statistics bias memory for individual items. *Psychological Science, 22*, 384-392. doi:10.1177/0956797610397956

Brady, T. F., & Tenenbaum, J. B. (2013). A probabilistic model of visual working memory: Incorporating higher order regularities into working memory capacity estimates. *Psychological Review, 120*, 85-109.

Della Sala, S., Logie, R., Trivelli, C., Cubelli, R., & Marchetti, C. (1998). Dissociation between recency and span: neuropsychological and experimental evidence. *Neuropsychology, 12*, 533-545.

Ecker, U. K. H., Lewandowsky, S., Oberauer, K., & Chee, A. E. H. (2010). The components of working memory updating: An experimental decomposition and individual differences. *Journal of Experimental Psychology-Learning Memory and Cognition, 36*(1), 170-189.

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review, 99*, 122-149.

Kane, M. J., & Engle, R. W. (2002). The role of prefrontal cortex in working-memory capacity, executive attention, and general fluid intelligence: An individual-differences perspective. *Psychonomic Bulletin and Review, 9*, 637-671.

Kessler, Y., & Oberauer, K. (2014). Working memory updating latency reflects the cost of switching between maintenance and updating modes of operation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*, 738-754.

Liefooghe, B., Barrouillet, P., Vandierendonck, A., & Camos, V. (2008). Working memory costs of task switching. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*, 478-494.

Logie, R. H. (in press). Scientific advance and theory integration in working memory: Commentary on Oberauer et al. (2018) Benchmarks for models of short-term and working memory. *Psychological Bulletin*.

Loncke, M., Desmet, T., Vandierendonck, A., & Hartsuiker, R. J. (2011). Executive control is shared between sentence processing and digit maintenance: Evidence from a strictly timed dual-task paradigm. *Journal of Cognitive Psychology, 23*(7), 886-911. doi:10.1080/20445911.2011.586625

Ma, W. J., Husain, M., & Bays, P. M. (2014). Changing concepts of working memory. *Nature Neuroscience Reviews, 17*, 347–356. doi:10.1038/nn.3655

Miyake, A., Carpenter, P. A., & Just, M. A. (1994). A capacity approach to syntactic comprehension disorders: Making normal adults perform like aphasic patients. *Cognitive Neuropsychology, 11*, 671-717.

Miyake, A., Carpenter, P. A., & Just, M. A. (1995). Reduced resources and specific impairments in normal and aphasic sentence comrehension. *Cognitive Neuropsychology, 12*, 651-679.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive Psychology, 41*, 49-100.

Nipher, F. E. (1878). On the distribution of errors in numbers written from memory. *Transactions of the Academy of Science of St. Louis, 3*, CCX-CCXI.

Oberauer, K., Lewandowsky, S., Awh, E., Brown, G. D. A., Cowan, N., Donkin, C., . . . Ward, G. (in press). Benchmarks for models of short-term and working memory. *Psychological Bulletin*.

Rac-Lubashevsky, R., & Kessler, Y. (2016). Dissociating working memory updating and automatic updating: The reference-back paradigm. *Journal of Experimental Psychology: Learning, Memory and Cognition, 42*, 951-969. doi:10.1037/xlm0000219

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review, 107*, 358-367.

Szmalec, A., Vandierendonck, A., & Kemps, E. (2005). Response selection involves executive control: Evidence from the selective interference paradigm. *Memory & Cognition, 33*, 531-541.

Vandierendonck, A. (in press). Working memory benchmarks: A missed opportunity. Comments on "Benchmarks for Models of Short Term and Working Memory". *Psychological Bulletin*.